Systematic Probing for Examining Evidence Based Tabular Reasoning

https://tabevidence.github.io/

Vivek Gupta[1*], Riyaz Bhat[2], Atreya Ghosal[3], Manish Shrivastava[3]
Maneesh Singh[2] , Vivek Srikumar[1]

[1]University of Utah; Verisk Analytics[2] ; IIIT Hyderabad[3]
*Bloomberg Ph.D. Fellow (2021-2022)

*on academic job market

Any *“evidence-based reasoning”* system should demonstrate *expected, predictable behavior* in *response* to *controlled changes to its inputs*.

Case Study on Tabular Inference

TAKEAWAY

Systematic target probing can highlights the limitation of tabular reasoning models

Such targeted probes are data efficient and work with minimal to no supervision

Probing can highlights several problems in tabular reasoning models:

(a) Use of annotation artifacts

(b) Use of incorrect evidence

(c) Overfitting on pre-trained knowledge

Check out the paper for details: https://tabprobe.github.io

TABULAR INFERENCE

- The **tabular natural language inference** problem is similar to standard NLI
- But here, the **premises are tabular data**
- Task: to decide whether given hypothesis is **true** (**entailment**), **false** (**contradiction**) or **undetermined** (**neutral**) given a premise table

Dressage

Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Contact	No
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

H1: Both men and women can complete in the contactless sport of Dressage → Entail

Check out InfoTabS (Gupta et al., 2020)
https://infotabs.github.io

EVIDENCE BASED REASONING

Both men and women can compete in the **contactless** sport of Dressage (**entail**)

But why **entail**? assume that model is

→ attention at **evidence** (relevant rows)
→ correct **reasoning** (logical inference)

black box problem

*"model **doesn't provide** the **highlighted evidence** and the **reasoning steps**"*

Dressage

Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Contact	No
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

CONTROLLED CHANGES

Both men and women can compete in the **contactless** sport of Dressage

remove the row with the key

contact

model prediction change

entail → ?

SYSTEMATIC PROBES

Dressage

Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Contact	No
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

EXPECTED MODEL RESPONSE

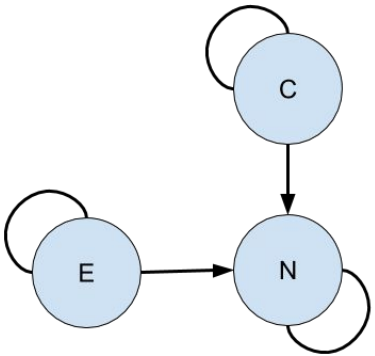
remove the key

contact from the table

Both men and women can compete in the **contactless** sport of Dressage

ideally prediction change

entail → neutral



Dressage

Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

SYSTEMATIC (‘CONTROLLED’) **PROBES** (CHANGES)

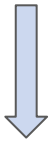
In our study, we define **three types** of systematic probes, as follow:

1. **Annotation Artifacts**:

   Can a model make inference about a hypothesis without a premise (a.k.a using artifacts)?

ANNOTATION ARTIFACTS

Can a model make inference about a hypothesis without a premise (a.k.a using artifacts)?



Modify the **hypothesis** in such a way that the inference **label is retained or flipped**

ANNOTATION ARTIFACTS

*Modify the following **types of expressions** to alter the **hypothesis***:

Named Entity	involving named entities
Numerical	related to numbers
Temporal	involving date and time
Quantification	related to introduction, deletion or modification of quantifiers such as most, many, every, etc.
Lexical	lexical semantics i.e. antonymy, synonymy, etc.
Negation	introduction or deletion of negation markers
Syntactic Alternations	leveraging syntactic structure
Subjective	adding subjective phrases/expressions

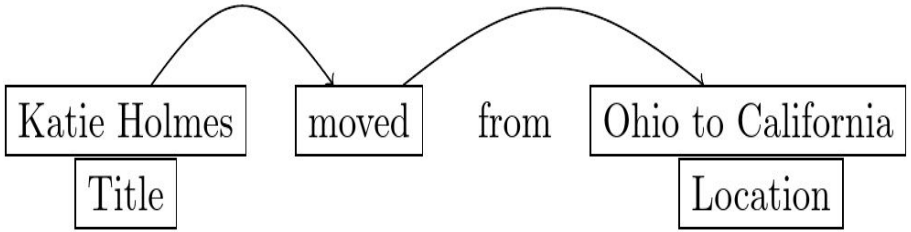
for others refer to the paper

EXAMPLE: ANNOTATION ARTIFACTS

Named Entity

H: Katie Holmes moved from **Ohio To California**.

Prediction: Entail



H': Katie Holmes moved from **South Africa** To California.

Prediction: Contradiction

title
Katie Holmes
Born
(1978-12-18) December 18, 1978 (age 40) Toledo, Ohio, U.S.
Birth name
Kate Noelle Holmes
Residence
Los Angeles, California, U.S.
Other names
Katherine Noelle Holmes
Occupation
Actress
Years active
1997-present
Spouse(s)
Tom Cruise (m. 2006 ; div. 2012)
Children
1

OUR FINDINGS

Model	Original $mean_{(stdev)}$	Label Preserved	Label Flipped
	Train Set (w/o NEUTRAL)		
Prem+Hypo	$99.44_{(0.06)}$	$92.98_{(0.20)}$	$53.92_{(0.28)}$
Hypo-Only	$96.39_{(0.13)}$	$70.23_{(0.35)}$	$19.23_{(0.27)}$
	α_1 Set (w/o NEUTRAL)		
Prem+Hypo	$68.94_{(0.76)}$	$69.56_{(0.77)}$	$51.48_{(0.86)}$
Hypo-Only	$63.52_{(0.75)}$	$60.27_{(0.85)}$	$31.02_{(0.63)}$

1. Model retain performance when label is preserve after perturbation

We tried two kinds of perturbations setting:

a) that preserve the label,

b) and those that flipped the label from entail to contradict and vice versa

OUR FINDINGS

Model	Original $mean_{(stdev)}$	Label Pre-served	Label Flipped
Train Set (w/o NEUTRAL)			
Prem+Hypo	$99.44_{(0.06)}$	$92.98_{(0.20)}$	$53.92_{(0.28)}$
Hypo-Only	$96.39_{(0.13)}$	$70.23_{(0.35)}$	$19.23_{(0.27)}$
α_1 Set (w/o NEUTRAL)			
Prem+Hypo	$68.94_{(0.76)}$	$69.56_{(0.77)}$	$51.48_{(0.86)}$
Hypo-Only	$63.52_{(0.75)}$	$60.27_{(0.85)}$	$31.02_{(0.63)}$

1. Model retain performance when label is preserve after perturbation

2. Substantial drop only for Hypothesis only baseline on training settings.

OUR FINDINGS

Model	Original $mean_{(stdev)}$	Label Pre-served	Label Flipped
Train Set (w/o NEUTRAL)			
Prem+Hypo	$99.44_{(0.06)}$	$92.98_{(0.20)}$	$53.92_{(0.28)}$
Hypo-Only	$96.39_{(0.13)}$	$70.23_{(0.35)}$	$19.23_{(0.27)}$
α_1 Set (w/o NEUTRAL)			
Prem+Hypo	$68.94_{(0.76)}$	$69.56_{(0.77)}$	$51.48_{(0.88)}$
Hypo-Only	$63.52_{(0.75)}$	$60.27_{(0.85)}$	$31.02_{(0.63)}$

1. Model retain performance when label is preserve after perturbation
2. Substantial drop only for Hypothesis only baseline on training settings.
3. Model overfit on hypothesis baseline in original setting

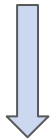
OUR FINDINGS

Model	Original $mean_{(stdev)}$	Label Pre-served	Label Flipped
	Train Set (w/o NEUTRAL)		
Prem+Hypo	$99.44_{(0.06)}$	$92.98_{(0.20)}$	$53.92_{(0.28)}$
Hypo-Only	$96.39_{(0.13)}$	$70.23_{(0.35)}$	$19.23_{(0.27)}$
	α_1 Set (w/o NEUTRAL)		
Prem+Hypo	$68.94_{(0.76)}$	$69.56_{(0.77)}$	$51.48_{(0.86)}$
Hypo-Only	$63.52_{(0.75)}$	$60.27_{(0.85)}$	$31.02_{(0.63)}$

1. Model retain performance when label is preserve after perturbation
2. Substantial drop only for Hypothesis only baseline on training settings.
3. Model overfit on hypothesis baseline in original setting.
4. Model performance dropped drastically when label is flipped after perturbation

FINDINGS: ANNOTATION ARTIFACTS

Can a model make inference about a hypothesis without a premise (a.k.a using artifacts)?



Yes, models largely rely on spurious correlation between hypothesis sentence and inference label.

for detailed results check the paper

Evidence Selection:

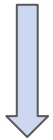
In our study, we define **three types** of systematic probes, as follow:

1. Annotation Artifacts:
    a. Can a model make inference about a hypothesis without a premise (a.k.a using artifacts)?

2. **Evidence Selection**:
    a. Is the model drawing inferences based on right evidence in the premise?

EVIDENCE SELECTION

Is the model drawing inferences based on right evidence in the premise?



Systematically **alter the premise table** via **simple operations** in order to **deterministically change the inference label.**

→ invalid label changes can be deterministically identified

POSSIBLE OPERATIONS

- Row Deletion
  - Automatic Probing: Random Row Deletion
  - Manual Probing: Require Row Annotation
    - Relevant Row Deletion
    - Irrelevant Row Deletion
- Row-Value Update
- New Row Insertion
- Row Perturbation

ANY ROW DELETION

Both men and women can play in the **contactless** sport of Dressage.

Prediction: Entail

Dressage

Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Contact	No
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

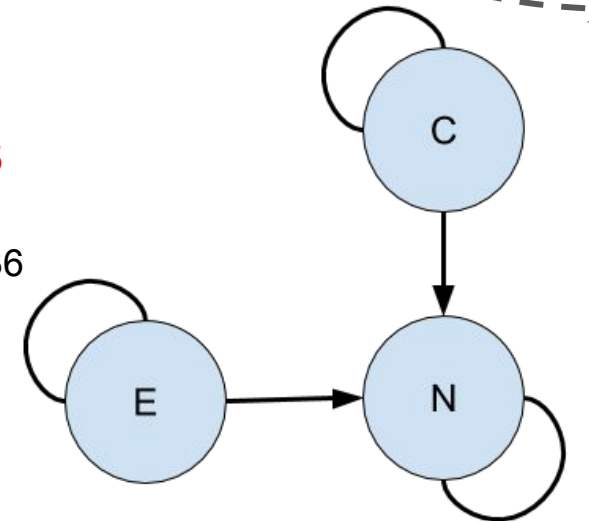
Random Row Deletion

remove the key **"*contact*"** from the table

Both men and women can play in the **contactless** sport of Dressage.

Prediction: **Neutral**

	E	N	C
E	3853	297	325
N	86	5341	131
C	212	278	5236



Dressage

Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

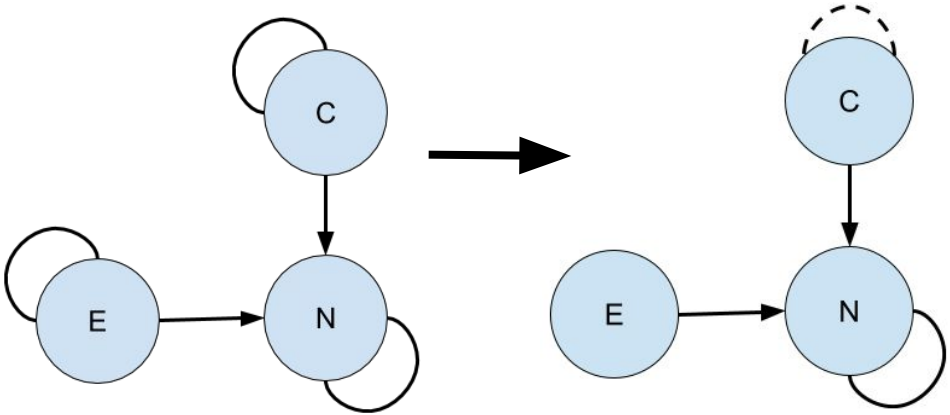
**RoBERTa Large* from Gupta, Vivek, et al. "INFOTABS: Inference on Tables as Semi-structured Data." *ACL* 2020.

Relevant Row Deletion

remove the key **"*contact*"** from the table

Both men and women can play in the **contactless** sport of Dressage.

Prediction: **Neutral**



Dressage

Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

***RoBERTa Large** from Gupta, Vivek, et al. "INFOTABS: Inference on Tables as Semi-structured Data." *ACL* 2020.

RELEVANT ROW DELETION

remove the key **"*contact*"** from the table

Both men and women can play in the **contactless** sport of Dressage.

Prediction: **Neutral**

	E	N	C
E	381	188	174
N	7	625	37
C	71	175	680

Dressage

Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

**RoBERTa Large* from Gupta, Vivek, et al. "INFOTABS: Inference on Tables as Semi-structured Data." *ACL* 2020.

ROW INSERTION

Insert row *"Arena size"* in the table

Both Men and Women can play in the **contactless** sport of Dressage.

Prediction : ??

Dressage

Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Contact	No
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Arena Size	20 by 60 m [66 by 197 ft]
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

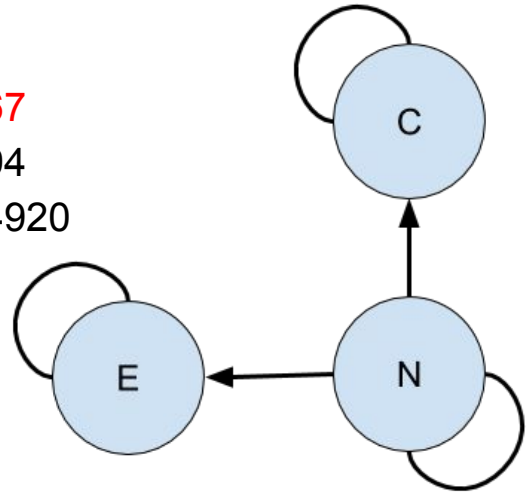
ROW INSERTION

insert row *"Arena size"* in the table

Both Men and Women can play in the **contactless** sport of Dressage.

Prediction : **Entail**

	E	N	C
E	10660	196	367
N	411	14168	294
C	695	355	14920

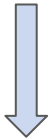


Dressage

Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Contact	No
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Arena Size	20 by 60 m [66 by 197 ft]
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

FINDING: EVIDENCE SELECTION

Is the model drawing inferences based on right evidence in the premise?



No, models does not look at correct evidence required for correct reasoning.

for detailed results check the paper

SYSTEMATIC ('CONTROL') PROBES (CHANGES):

In our study, we define **three types** of systematic probes, as follow:

1. Annotation Artifacts:

   Can a model make inference about a hypothesis without a premise (a.k.a using artifacts)?

2. Evidence Selection:

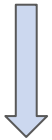
   Is the model drawing inferences based on right evidence in the premise?

3. **Counterfactual Instances:**

   How will the model react if the primary evidence is counterfactual to pre-trained data?

Counterfactual Instances

How will the model react if the primary evidence is counterfactual to pre-trained data?



Update the premise table to include counterfactual data in order to retain or change the inference label.

Example: Counterfactual Instances

Dressage

Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Contact	No
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

Both men and women can complete in the **contactless** sport of Dressage → Entail

Dressage

Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Contact	Yes
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

Both men and women can complete in the **contactless** sport of Dressage → Contradict

Example: Counterfactual Instances

Dressage

Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Contact	No
Team members	Individual and team at international levels
Mixed gender	Yes
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

Dressage

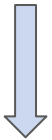
Highest governing body	International Federation for Equestrian Sports (FEI)
Characteristics	
Contact	No
Team members	Individual and team at international levels
Mixed gender	No
Equipment	Horse, appropriate horse tack
Venue	Arena, indoor or outdoor
Presence	
Country or region	Worldwide
Olympic	1912
Paralympic	1996

Both men and women can complete in the contactless sport of Dressage → Entail

Both men and women can complete in the contactless sport of Dressage → Contradict

FINDINGS: ANNOTATION ARTIFACTS

How will the model react if the primary evidence is counterfactual to pre-trained data?



Model relies on information from pre-trained language models rather than tabular evidence for making prediction

for detailed results check the paper

Main Findings

- **Annotation Artifacts:** Model largely rely on spurious correlation between hypothesis and inference label.
- **Evidence Selection:** The model does not look at correct evidence required for correct reasoning.
- **Counterfactual Instances:** Model relies on information from pre-trained language model rather than tabular evidence.
- **Inoculation Study:** Changes in the data distribution during training have a negative impact on model performance.

TAKEAWAY

Systematic target probing can highlight the limitation of tabular reasoning models

Such targeted probes are data efficient and work with minimal to no supervision

Probing can highlights problems in tabular reasoning models:

(a) Use of annotation artifacts

(b) Use of incorrect evidence

(c) Overfitting on pre-trained knowledge

Check out the paper for details: https://tabprobe.github.io